Punny Captions: Witty Wordplay in Image Descriptions

Arjun Chandrasekaran[1] **Devi Parikh**[1,2] **Mohit Bansal**[3]
[1]Georgia Institute of Technology [2]Facebook AI Research [3]UNC Chapel Hill
{carjun, parikh}@gatech.edu mbansal@cs.unc.edu

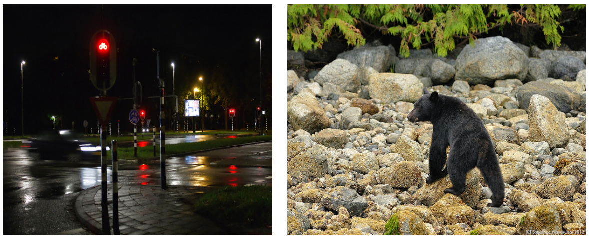
Abstract

Wit is a form of rich interaction that is often grounded in a specific situation (e.g., a comment in response to an event). In this work, we attempt to build computational models that can produce witty descriptions for a given image. Inspired by a cognitive account of humor appreciation, we employ linguistic wordplay, specifically puns, in image descriptions. We develop two approaches which involve retrieving witty descriptions for a given image from a large corpus of sentences, or generating them via an encoder-decoder neural network architecture. We compare our approach against meaningful baseline approaches via human studies and show substantial improvements. We find that when a human is subject to similar constraints as the model regarding word usage and style, people vote the image descriptions generated by our model to be slightly wittier than human-written witty descriptions. Unsurprisingly, humans are almost always wittier than the model when they are free to choose the vocabulary, style, etc.

1 Introduction

"*Wit is the sudden marriage of ideas which before their union were not perceived to have any relation.*" – Mark Twain. Witty remarks are often contextual, i.e., grounded in a specific situation. Developing computational models that can emulate rich forms of interaction like contextual humor, is a crucial step towards making human-AI interaction more natural and more engaging (Yu et al., 2016). E.g., witty chatbots could help relieve stress and increase user engagement by being more personable and human-like. Bots could automatically post witty comments (or suggest witty responses) on social media, chat, or messaging.

The absence of large scale corpora of witty captions and the prohibitive cost of collecting such a dataset (being witty is harder than just describing an image) makes the problem of producing contextually witty image descriptions challenging.



(a) **Generated**: a poll (pole) on a city street at night.
Retrieved: the light knight (night) chuckled.
Human: the knight (night) in shining armor drove away.

(b) **Generated**: a bare (bear) black bear walking through a forest.
Retrieved: another reporter is standing in a bare (bear) brown field.
Human: the bear killed the lion with its bare (bear) hands.

Figure 1: Sample images and witty descriptions from 2 models, and a human. The words inside '()' (e.g., pole and bear) are the puns associated with the image, i.e., the source of the unexpected puns used in the caption (e.g., poll and bare).

In this work, we attempt to tackle the challenging task of producing witty (pun-based) remarks for a given (possibly boring) image. Our approach is inspired by a two-stage cognitive account of humor appreciation (Suls, 1972) which states that a perceiver experiences humor when a stimulus such as a joke, captioned cartoon, etc., causes an *incongruity*, which is shortly followed by *resolution*.

We introduce an incongruity in the perceiver's mind while describing an image by using an unexpected word that is phonetically similar (pun) to a concept related to the image. E.g., in Fig. 1b, the expectations of a perceiver regarding the image (bear, stones, etc.) is momentarily disconfirmed by the (phonetically similar) word 'bare'. This incongruity is resolved when the perceiver parses the entire image description. The incongruity followed by resolution can be perceived to be witty.[1]

[1] Indeed, a perceiver may fail to appreciate wit if the pro-

Proceedings of NAACL-HLT 2018, pages 770–775
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

We build two computational models based on this approach to produce witty descriptions for an image. First, a model that retrieves sentences containing a pun that are relevant to the image from a large corpus of stories (Zhu et al., 2015). Second, a model that generates witty descriptions for an image using a modified inference procedure during image captioning which includes the specified pun word in the description.

Our paper makes the following contributions: To the best of our knowledge, this is the first work that tackles the challenging problem of producing a witty natural language remark in an everyday (boring) context. We present two novel models to produce witty (pun-based) captions for a novel (likely boring) image. Our models rely on linguistic wordplay. They use an unexpected pun in an image description during inference/retrieval. Thus, they do not require to be trained with witty captions. Humans vote the descriptions from the top-ranked *generated* captions 'wittier' than three baseline approaches. Moreover, in a Turing test-style evaluation, our model's best image description is found to be wittier than a witty human-written caption[2] 55% of the time when the human is subject to the same constraints as the machine regarding word usage and style.

2 Related Work

Humor theory. General Theory of Verbal Humor (Attardo and Raskin, 1991) characterizes linguistic stimuli that induce humor but implementing computational models of it requires severely restricting its assumptions (Binsted, 1996).

Puns. Zwicky and Zwicky (1986) classify puns as *perfect* (pronounced exactly the same) or *imperfect* (pronounced differently). Similarly, Pepicello and Green (1984) categorize riddles based on the linguistic ambiguity that they exploit – phonological, morphological or syntactic. Jaech et al. (2016) learn phone-edit distances to predict the counterpart, given a pun by drawing from automatic speech recognition techniques. In contrast, we augment a web-scraped list of puns using an existing model of pronunciation similarity.

Generating textual humor. JAPE (Binsted and Ritchie, 1997) also uses phonological ambiguity to generate pun-based riddles. While our task involves producing free-form responses to a novel stimulus, JAPE produces stand-alone "canned" jokes. HAHAcronym (Stock and Strapparava, 2005) generates a funny expansion of a given acronym. Unlike our work, HAHAcronym operates on text, and is limited to producing sets of words. Petrovic and Matthews (2013) develop an unsupervised model that produces jokes of the form, "*I like my X like I like my Y, Z*" .

Generating multi-modal humor. Wang and Wen (2015) predict a meme's text based on a given funny image. Similarly, Shahaf et al. (2015) and Radev et al. (2015) learn to rank cartoon captions based on their funniness. Unlike typical, boring images in our task, memes and cartoons are images that are already funny or atypical. E.g., "LOL-cats" (funny cat photos), "Bieber-memes" (modified pictures of Justin Bieber), cartoons with talking animals, etc. Chandrasekaran et al. (2016) alter an abstract scene to make it more funny. In comparison, our task is to generate witty natural language remarks for a novel image.

Poetry generation. Although our tasks are different, our generation approach is conceptually similar to Ghazvininejad et al. (2016) who produce poetry, given a topic. While they also generate and score a set of candidates, their approach involves many more constraints and utilizes a finite state acceptor unlike our approach which enforces constraints during beam search of the RNN decoder.

3 Approach

Extracting tags. The first step in producing a contextually witty remark is to identify concepts that are relevant to the context (image). At times, these concepts are directly available as e.g., tags posted on social media. We consider the general case where such tags are unavailable, and automatically extract tags associated with an image.

We extract the top-5 object categories predicted by a state-of-the-art Inception-ResNet-v2 model (Szegedy et al., 2017) trained for image classification on ImageNet (Deng et al., 2009). We also consider the words from a (boring) image description (generated from Vinyals et al. (2016)). We combine the classifier object labels and words from the caption (ignoring stopwords) to produce a set of tags associated with an image, as shown in Fig. 2. We then identify concepts from this collection that can potentially induce wit.

Identifying puns. We attempt to induce an incongruity by using a pun in the image description. We identify candidate words for linguistic wordplay

cess of 'solving' (resolution) is trivial (the joke is obvious) or too complex (they do not 'get' the joke).

[2] This data is available on the author's webpage.

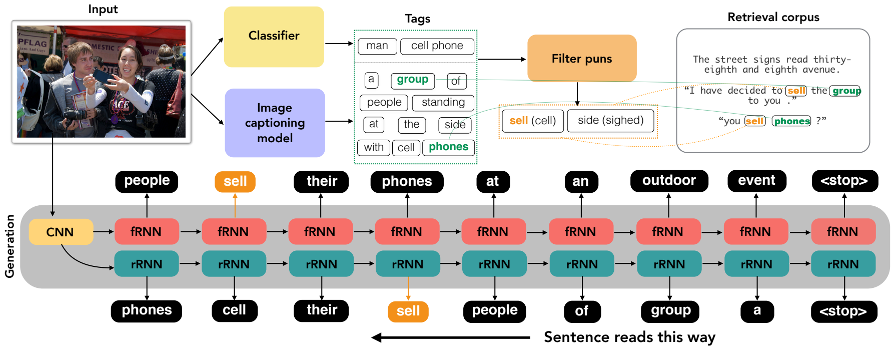


Figure 2: Our models for generating and retrieving image descriptions containing a pun (see Sec. 3).

by comparing image tags against a list of puns.

We construct the list of puns by mining the web for differently spelled words that sound exactly the same (heterographic homophones). We increase coverage by also considering pairs of words with 0 edit-distance, according to a metric based on fine-grained articulatory representations (AR) of word pronunciations (Jyothi and Livescu, 2014). Our list of puns has a total of 1067 unique words (931 from the web and 136 from the AR-based model).

The pun list yields a set of puns that are associated with a given image and their phonologically identical counterparts, which together form the *pun vocabulary* for the image. We evaluate our approach on the subset of images that have non-empty pun vocabularies (about 2 in 5 images).

Generating punny image captions. We introduce an incongruity by forcing a vanilla image captioning model (Vinyals et al., 2016) to decode a *phonological counterpart* of a pun word associated with the image, at a specific time-step during inference (e.g., ‘sell’ or ‘sighed’, showed in orange in Fig. 2). We achieve this by limiting the vocabulary of the decoder at that time-step to only contain counterparts of image-puns. In following time-steps, the decoder generates new words conditioned on all previously decoded words. Thus, the decoder attempts to generate sentences that flow well based on previously uttered words.

We train two models that decode an image description in forward (start to end) and reverse (end to start) directions, depicted as ‘fRNN’ and ‘rRNN’ in Fig. 2 respectively. The fRNN can decode words after accounting for the incongruity that occurs early in the sentence and the rRNN is able to decode the early words in the sentence after accounting for the incongruity that can occur later. The forward RNN and reverse RNN generate sentences in which the pun appears in each of the first T and last T positions, respectively.[3]

Retrieving punny image captions. As an alternative to our approach of *generating* witty remarks for the given image, we also attempt to leverage natural, human-written sentences which are relevant (yet unexpected) in the given context. Concretely, we retrieve natural language sentences[4] from a combination of the Book Corpus (Zhu et al., 2015) and corpora from the NLTK toolkit (Loper and Bird, 2002). The retrieved sentences each (a) contains an incongruity (pun) whose counterpart is associated with the image, and (b) has support in the image (contains an image tag). This yields a pool of candidate captions that are perfectly grammatical, a little unexpected, and somewhat relevant to the image (see Sec. 4).

Ranking. We rank captions in the candidate pools from both generation and retrieval models, according to their log-probability score under the image captioning model. We observe that the higher-ranked descriptions are more relevant to the image and grammatically correct. We then perform non-maximal suppression, i.e., eliminate captions that are similar[5] to a higher-ranked caption to reduce the pool to a smaller, more diverse set. We report results on the 3 top-ranked captions. We describe the effect of design choices in the supplementary.

4 Results

Data. We evaluate witty captions from our approach via human studies. 100 random images (having associated puns) are sampled from the val-

[3] For an image, we choose $T = \{1, 2, ..., 5\}$ and beam size $= 6$ for each decoder. This generates a pool of 5 (T) $*$ 6 (beam size) $*$ 2 (forward + reverse decoder) $=$ 60 candidates.

[4] To prevent the context of the sentence from distracting the perceiver, we consider sentences with < 15 words. Overall, we are left with a corpus of about 13.5 million sentences.

[5] Two sentences are similar if the cosine similarity between the average of the Word2Vec (Mikolov et al., 2013) representations of words in each sentence is ≥ 0.8.

idation set of COCO (Lin et al., 2014).

Baselines. We compare the wittiness of descriptions generated by our model against 3 qualitatively different baselines, and a human-written witty description of an image. Each of these evaluates a different component of our approach. **Regular inference** generates a fluent caption that is relevant to the image but is not attempting to be witty. **Witty mismatch** is a human-written witty caption, but for a different image from the one being evaluated. This baseline results in a caption that is intended to be witty, but does not attempt to be relevant to the image. **Ambiguous** is a 'punny' caption where a pun word in the boring (regular) caption is replaced by its counterpart. This caption is likely to contain content that is relevant to the image, *and* it contains a pun. However, the pun is not being used in a fluent manner.

We evaluate the **image-relevance** of the top witty caption by comparing against a boring machine caption and a random caption (see supplementary).

Evaluation annotations. Our task is to generate captions that a layperson might find witty. To evaluate performance on this task, we ask people on Amazon Mechanical Turk (AMT) to vote for the wittier among the given pair of captions for an image. We collect annotations from 9 unique workers for each relative choice and take the majority vote as ground-truth. For each image, we compare each of the generated 3 top-ranked and 1 low-ranked caption against 3 baseline captions and 1 human-written witty caption.[6]

Constrained human-written witty captions. We evaluate the ability of humans and automatic methods to use the given context *and pun words* to produce a caption that is perceived as witty. We ask subjects on AMT to describe a given image in a witty manner. To prevent observable *structural* differences between machine and human-written captions, we ensure consistent pun vocabulary (utilization of pre-specified puns for a given image). We also ask people to avoid first person accounts or quote characters in the image.

Metric. 3, we report performance of the generation approach using the Recall@K metric. For $K = 1, 2, 3$, we plot the percentage of images for which at least one of the K 'best' descriptions from our model outperformed another approach.

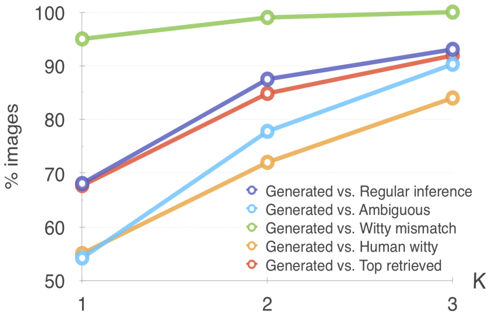


Figure 3: Wittiness of top-3 generated captions vs. other approaches. y-axis measures the % images for which at least one of K captions from our approach is rated wittier than other approaches. Recall steadily increases with the number of generated captions (K).

Generated captions vs. baselines. As we see in Fig. 3, the top generated image description (top-1G) is perceived as wittier compared to all baseline approaches more often than not (the vote is $>50\%$ at $K = 1$). We observe that as K increases, the recall steadily increases, i.e., when we consider the top K generated captions, increasingly often, humans find at least one of them to be wittier than captions produced by baseline approaches. People find the top-1G for a given image to be wittier than mismatched human-written image captions, about 95% of the time. The top-1G is also wittier than a naive approach that introduces ambiguity about 54.2% of the time. When compared to a typical, boring caption, the generated captions are wittier 68% of the time. Further, in a head-to-head comparison, the generated captions are wittier than the retrieved captions 67.7% of the time. We also validate our choice of ranking captions based on the image captioning model score. We observe that a 'bad' caption, i.e., one ranked lower by our model, is significantly less witty than the top 3 output captions.

Surprisingly, when the human is constrained to use the same words and style as the model, the generated descriptions from the model are found to be wittier for 55% of the images. Note that in a Turing test, a machine would equal human performance at 50%[7]. This led us to speculate if the con-

[6] This results in a total of 4 (captions) $*2$ (generation + retrieval) $*4$ (baselines + human caption) = 32 comparisons of our approach against baselines. We also compare the wittiness of the 4 generated captions against the 4 retrieved captions (see supplementary) for an image (16 comparisons). In total, we perform 48 comparisons per image, for 100 images.

[7] Recall that this compares how a witty description is constructed, given the image *and* specific pun words. A Turing test-style evaluation that compares the overall wittiness of a machine and a human would refrain from constraining the human in any way.

(a) **Generated**: a bored (board) bench sits in front of a window.
Retrieved: Wedge sits on the bench opposite Berry, bored (board).
Human: could you please make your pleas (please)!

(b) **Generated**: a loop (loupe) of flowers in a glass vase.
Retrieved: the flour (flower) inside teemed with worms.
Human: piece required for peace (piece).

(c) **Generated**: a woman sell (cell) her cell phone in a city.
Retrieved: Wright (right) slammed down the phone.
Human: a woman sighed (side) as she regretted the sell.

(d) **Generated**: a bear that is bare (bear) in the water.
Retrieved: water glistened off her bare (bear) breast.
Human: you won't hear a creak (creek) when the bear is feasting.

(e) **Generated**: a loop (loupe) of scissors and a pair of scissors.
Retrieved: i continued slicing my pear (pair) on the cutting board.
Human: the scissors were near, but not clothes (close).

(f) **Generated**: a female tennis player caught (court) in mid swing.
Retrieved: i caught (court) thieves on the roof top.
Human: the man made a loud bawl (ball) when she threw the ball.

(g) **Generated**: a bored (board) living room with a large window.
Retrieved: anya sat on the couch, feeling bored (board).
Human: the sealing (ceiling) on the envelope resembled that in the ceiling.

(h) **Generated**: a parking meter with rode (road) in the background.
Retrieved: smoke speaker sighed (side).
Human: a nitting of color didn't make the poll (pole) less black.

Figure 4: The top row contains selected examples of human-written witty captions, and witty captions generated and retrieved from our models. The examples in the bottom row are randomly picked.

straints placed on language and style might be restricting people's ability to be witty. We confirmed this by evaluating free-form human captions.

Free-form Human-written Witty Captions. We ask people on AMT to describe an image (using any vocabulary) in a manner that would be perceived as funny. As expected, when compared against automatic captions from our approach, human evaluators find free-form human captions to be wittier about 90% of the time compared to 45% in the case of constrained human witty captions. Clearly, human-level creative language with unconstrained sentence length, style, choice of puns, etc., makes a significant difference in the wittiness of a description. In contrast, our automatic approach is constrained by caption-like language, length, and a word-based pun list. Training models to intelligently navigate this creative freedom is an exciting open challenge.

Qualitative analysis. The generated witty captions exhibit interesting features like alliteration ('a bare black bear ...') in Fig. 1b and 4c. At times, both the original pun (pole) and its counterpart (poll) make sense for the image (Fig. 1a). Occasionally, a pun is naively replaced by its counterpart (Fig. 4a) or rare puns are used (Fig. 4b). On the other hand, some descriptions (Fig. 4e and 4h) that are forced to utilize puns do not make sense. See supplementary for analysis of retrieval model.

5 Conclusion

We presented novel computational models inspired by cognitive accounts to address the challenging task of producing contextually witty descriptions for a given image. We evaluate the models via human-studies, in which they significantly outperform meaningful baseline approaches.

Acknowledgements

We thank Shubham Toshniwal for his advice regarding the automatic speech recognition model. This work was supported in part by: a NSF CAREER award, ONR YIP award, ONR Grant N00014-14-12713, PGA Family Foundation award, Google FRA, Amazon ARA, DARPA XAI grant to DP and NVIDIA GPU donations, Google FRA, IBM Faculty Award, and Bloomberg Data Science Research Grant to MB.

References

Salvatore Attardo and Victor Raskin. 1991. Script theory revis (it) ed: Joke similarity and joke representation model. *Humor-International Journal of Humor Research* 4(3-4):293–348.

Kim Binsted. 1996. Machine humour: An implemented model of puns. *PhD Thesis, University of Edinburgh* .

Kim Binsted and Graeme Ritchie. 1997. Computational rules for generating punning riddles. *Humor: International Journal of Humor Research* .

Arjun Chandrasekaran, Ashwin Kalyan, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2016. We are humor beings: Understanding and predicting visual humor. In *CVPR*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, pages 248–255.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *EMNLP*.

Aaron Jaech, Rik Koncel-Kedziorski, and Mari Ostendorf. 2016. Phonological pun-derstanding. In *HLT-NAACL*.

Preethi Jyothi and Karen Livescu. 2014. Revisiting word neighborhoods for speech recognition. *ACL 2014* page 1.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, pages 740–755.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. Association for Computational Linguistics, ETMTNLP '02.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

William J Pepicello and Thomas A Green. 1984. *Language of riddles: new perspectives*. The Ohio State University Press.

Sasa Petrovic and David Matthews. 2013. Unsupervised joke generation from big data. In *ACL*.

Dragomir Radev, Amanda Stent, Joel Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, et al. 2015. Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. *arXiv preprint arXiv:1506.08126* .

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pages 1065–1074.

Oliviero Stock and Carlo Strapparava. 2005. HAHAcronym: A computational humor system. In *ACL*.

Jerry M Suls. 1972. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. *The Psychology of Humor: Theoretical Perspectives and Empirical Issues* .

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*. pages 4278–4284.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence* .

William Yang Wang and Miaomiao Wen. 2015. I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. In *NAACL*.

Zhou Yu, Leah Nicolich-Henkin, Alan W Black, and Alex I Rudnicky. 2016. A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. page 55.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*. pages 19–27.

Arnold Zwicky and Elizabeth Zwicky. 1986. Imperfect puns, markedness, and phonological similarity: With fronds like these, who needs anemones. *Folia Linguistica* 20(3-4):493–503.